COMPOSITE ESTIMATION IN NATIONAL AND STATE SURVEYS

John Burke, Leyla Mohadjer, Jim Green, and Joseph Waksberg, Westat, Inc.
Irwin S. Kirsch, Educational Testing Service, Andrew Kolstad, NCES
John Burke, Westat Inc., 1650 Research Boulevard, Rockville, Maryland 20850

Key words: composite estimation, sampling weights

1. SAMPLE DESIGN

The National Adult Literacy Survey (NALS) was conducted in 1992 with a nationally representative sample of some 13,600 adults aged 16 and older, each of whom was asked to provide personal and background information and to complete a booklet of literacy tasks. Black and Hispanic households were oversampled to ensure reliable estimates of literacy proficiencies and to permit analyses of the performance of these subpopulations.

To give states an opportunity to explore the skill levels of their populations, each of the 50 states was invited to participate in a concurrent assessment. While many states expressed an interest, 11 (California, Louisiana, Pennsylvania, Illinois, New Jersey, Texas, Indiana, New York, Washington, Iowa and Ohio) elected to participate in the State Adult Literacy Survey. Approximately 1,000 adults aged 16 to 64 were surveyed in each of the states.

This paper presents the methods used to composite the national and state data in NALS and evaluates the compositing factors using actual literacy scores.

The national and state household components were based on a four-stage, stratified area sample with the following stages: (1) the selection of primary sampling units (PSUs) consisting of counties or groups of counties, (2) the selection of segments consisting of census blocks or groups of blocks, (3) the selection of households, and (4) the selection of age-eligible individuals. A single area sample was drawn for the national component and 11 additional independent state-level area samples were drawn for the state component. The national and state samples differed in two important respects. In the national sample, blacks and Hispanics were sampled at a higher rate than the remainder of the population to increase their representation in the sample, whereas the state samples used no oversampling. Also, the target population for the national sample consisted of adults 16 years of age or older, whereas the target population for the state samples consisted of adults aged 16 to 64.

2. OVERVIEW OF WEIGHTING

Base weights were computed as the reciprocal of the product of the probabilities of selection at each stage of sampling. Before compositing the national and state samples, the base weights for each sample were poststratified separately to known population totals.

After compositing the national and state samples, the final sampling weights were computed by raking the composited weights to known population totals. The variables used to construct raking classes for NALS were age, race/ethnicity (blacks, non-black Hispanics, and others), sex, education, and geographic indicators, i.e., metropolitan statistical area (MSA) vs. non-MSA for the 11 states and census region for the remainder of the United States.

The 1990 population totals used for raking were adjusted to account for census undercoverage. The undercoverage rates, based on information provided by the U.S. Bureau of the Census, were applied separately by age, race/ethnicity, sex, and region of the country .

3. DETAILS OF COMPOSITING

Composite weights were developed so that NALS data could be used to produce both state and national statistics. The composite estimator for the national/state sample is given by

$$\hat{Y}_{ik} = \beta_{ik}\hat{Y}_{(st)ik} + (1-\beta_{ik})\hat{Y}_{(nt)ik} \qquad (1)$$

where

$\hat{Y}_{ik}$ = the composite estimator for variable Y in state i for subgroup k;

β_{ik} = the compositing factor in state i for subgroup k;

$\hat{Y}_{(st)ik}$ = the estimate of Y coming from state i for subgroup k; and

$\hat{Y}_{(nt)ik}$ = the estimate of Y coming from the national sample in state i for subgroup k.

For statistic $\hat{Y}_{ik}$, the optimal compositing factor for state i and subgroup k, is

$$\beta_{ik} = \frac{V(\hat{Y}_{(nt)ik})}{V(\hat{Y}_{(nt)ik}) + V(\hat{Y}_{(st)ik})} \qquad (2)$$

where

$V(\hat{Y}_{(nt)ik})$ = the variance of the estimate of Y coming from the national sample in state i for subgroup k; and

$V(\hat{Y}_{(st)ik})$ = the variance of the estimate of Y coming from the state sample in state i for subgroup k.

A different optimal value of β_{ik} might be found for each statistic of interest. Under simple random sampling, the variance of the estimator is inversely proportional to the sample size, and the expression for β_{ik} simplifies to the following:

$$\beta_{ik} = \frac{n_{(st)ik}}{n_{(st)ik} + n_{(nt)ik}}$$

where

$n_{(st)ik}$ = the number of respondents aged 16 to 64 in the state sample; and

$n_{(nt)ik}$ = the number of respondents aged 16 to 64 in the national sample.

Because of the complexity of the NALS sample design, it was useful to think of deriving β_{ik} in terms of the effective sample size, i.e., the actual sample size divided by the design effect. Three aspects of the NALS design tended to influence the design effect and thereby reduce the effective sample size overall: clustering, stratification, and differential sampling rates used for blacks and Hispanics.

To best reflect the influence of these design aspects on the effective sample size, distinct compositing factors were derived for up to four subsets of data in each participating state. Those subsets were defined according to (1) whether or not the data came from a PSU chosen with certainty for the national sample and (2) whether or not the respondent was black or Hispanic.

For data collected in PSUs selected with certainty for both the national and state samples, the effective sample size was estimated as

$$n_{ijk_{eff}} = \frac{n_{ijk}}{1+\left(\bar{n}_{ijk}-1\right)\rho_1 + V^2_{w_{ijk}}}$$

where

i = a participating state;

j = national or state sample;

k = minority (black or Hispanic) or non minority;

n_{ijk} = total number of respondents aged 16 to 64;

$\bar{n}_{ijk}$ = mean number of respondents per segment;

ρ_1 = 0.042, the intraclass correlation within segment, assumed to be equal to the CPS average and to be constant across states; and

$V^2_{w_{ijk}}$ = the relvariance of the weights.

For data collected in other than the certainty PSUs included in the national sample, the effective sample size was estimated as

$$n_{ijk_{eff}} = \frac{n_{ijk}}{1+\left(\bar{n}_{ijk}-1\right)\rho_1 + \left(\bar{m}_{ijk}-1\right)\rho_2 P_{ijk} F_{ij} + V^2_{w_{ijk}}}$$

where

i = a participating state;

j = national or state sample;

k = minority (black or Hispanic) or non minority;

n_{ijk} = total number of respondents aged 16 to 64;

$\bar{n}_{ijk}$ = mean number of respondents per segment;

ρ_1 = 0.042, the intraclass correlation within segment, assumed to be equal to the CPS average and to be constant across states;

$\bar{m}_{ijk}$ = mean number of respondents per PSU;

ρ_2 = 0.00075, the intraclass correlation within PSU, assumed to be equal to the CPS average and to be constant across states;

P_{ijk} = the proportion of respondents in noncertainty PSUs;

F_{ij} = a design-effect-like factor descriptive of the relative inefficiency of the national PSU sample design for making state estimates; and

$V^2_{w_{ijk}}$ = the relvariance of the weights.

Then an estimate of the optimal composite factor for state i was given by

$$\beta_{i(State)k} = \frac{n_{i(State)k_{eff}}}{n_{i(State)k_{eff}} + n_{i(National)k_{eff}}}$$

and

$$\beta_{i(National)k} = 1 - \beta_{i(State)k} = \frac{n_{i(National)k_{eff}}}{n_{i(State)k_{eff}} + n_{i(National)k_{eff}}}$$

4. EVALUATING THE EFFECTIVENESS OF NALS COMPOSITING

The main objective of compositing the national and state samples was to improve the precision of the estimates. The composite estimation did improve the statistics coming from the 11 state samples. It also improved the precision of statistics for the nation, but the relative gain was lower than for the 11 states. Table 1 shows the percent decrease in variance for national prose proficiency statistics after compositing the national and state data. The table also presents the percent increase in the sample size after compositing the data. The general pattern indicates that the variances were decreased as a result of compositing but at a much lower rate than the increase in sample sizes. This is not a surprising outcome because the additional sample size came from 11 states that made up about one-half of the total U.S. population. In some cases, the percent decrease is a negative number, indicating that variances were increased as a result of compositing. It should be noted, however, that the variances of some

Table 1. NALS Compositing Analysis: Percent Change in Sample Size and Variance After Compositing for Average Prose Proficiency and Literacy Levels by Total Population, Gender, Census Region, Race/Ethnicity, Education Level, Age, and Country of Birth

		Percent Decrease in Variance After Compositing					
Demographic Subpopulations	Percent Increase in Sample Size	Level 1 225 or lower	Level 2 226 to 275	Level 3 276 to 325	Level 4 326 to 375	Level 5 376 or higher	Overall Proficiency
Total Population							
Total	83.5	38.5	17.0	-11.9	12.1	50.5	18.7
Gender							
Male	87.7	31.6	-3.6	-38.0	42.0	35.7	16.6
Female	80.6	38.0	9.2	52.0	15.3	40.0	20.0
Census Region							
Northeast	113.9	61.1	11.9	58.5	53.2	38.0	62.4
Midwest	137.2	39.6	32.3	16.0	19.6	57.7	44.2
South	43.6	-19.3	10.5	11.7	3.2	9.2	-26.9
West	73.0	-1.2	-2.1	-53.9	39.4	61.7	-3.7
Race/Ethnicity							
Black	41.4	16.4	-31.9	13.5	4.8	36.5	11.5
Hispanic	46.6	44.9	43.0	27.2	40.7	26.1	35.0
Other	108.1	50.9	32.5	0.1	22.8	52.3	36.7
Education Level							
No HS degree	55.9	22.4	46.3	15.8	15.7	-5.1	19.7
HS degree	88.0	43.4	22.8	4.4	-1.2	44.9	25.5
Some college	95.6	30.5	31.1	18.9	27.9	37.5	29.9
College graduate	101.7	20.0	10.9	-2.8	25.2	57.1	25.3
Age							
16 to 24 years							
25 to 44 years	88.0	18.5	35.4	18.9	21.2	23.4	38.7
45 to 64 years	97.7	54.2	23.7	1.2	19.0	39.0	29.1
	108.6	34.6	27.7	15.4	18.6	39.6	2.6
Country of Birth							
Not USA	63.8	23.1	36.5	11.5	33.7	26.8	6.2
USA	86.1	26.5	18.3	-9.8	17.6	49.5	6.3

of the items in the table are quite small, making the ratio (the estimate of the percent decrease in variance) very unstable. For example, the estimated variances of level 3 prose literacy scores for males are 0.000105 and 0.000145 before and after compositing, respectively. The difference between the two estimates is trivial, even though the table shows a 38% increase in the variance. Another factor that should be considered when studying this table is that the entries are estimates themselves and are subject to variation.

5. USING NALS LITERACY SCORES TO ESTIMATE NEW COMPOSITING FACTORS

The standard theoretical foundation of composite estimation requires a knowledge of variances of the statistics of interest. This information is necessary to produce the parameters used to combine data from various surveys in a way that minimizes the variances of the composite estimates. However, the composite weighting had to be completed before literacy score data were available for NALS. After the literacy data became available, new compositing factors were computed for a selected set of statistics. This chapter presents the new compositing factors and describes the methods used to estimate them.

The estimation of the new factors required computation of components of variance for a set of statistics chosen from the NALS data set. Estimates of variances, design effects, and compositing factors were computed from the NALS data for (1) mean proficiency scores and (2) the percentage of persons scoring at each of five literacy levels: 225 or lower, 226 to 275, 276 to 325, 326 to 375 and 376 or higher.

Estimates were computed for the following population totals: total population, sex, Census region, race/ethnicity (Hispanic; black, non-Hispanic; and other), education (less than high school diploma, high school diploma, some college, and college graduate), age (16 to 24, 25 to 44, and 45 to 64); and country of birth (born in or outside of the United States).

For a given population total, the usual unbiased weighted estimator is defined by

$$y' = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where

y' = the unbiased estimate of the population total Y

n = the sample size;

y_i = the reported value of the characteristic for the i$^{\text{th}}$ person in the sample; and

π_i = the probability of selection for the i$^{\text{th}}$ respondent.

The variance of the estimate can be expressed in the following summarized form:

$$\sigma^2(y') = \sigma_B^2(y') + \sigma_W^2(y') \qquad (3)$$

where

$\sigma^2(y')$ = the total variance of the estimate;

$\sigma_B^2(y')$ = the between-PSU component of variance; and

$\sigma_W^2(y')$ = the within-PSU variance.

The between-PSU component of variance reflects the contribution to variance that results from the sampling of PSUs. The within-PSU component reflects variability arising from several sources, including variance resulting from the selection of segments within PSUs, the selection of households within segments, and the selection of more than one person per household. This component also reflects the additional variability arising from the variation in weights due to the oversampling of blacks and Hispanics in segments with high concentrations of these minorities and the subsampling of persons within households.

Estimates of the components of variance were computed using the jackknife replication method. Under this approach, a set of replicates is formed where each replicate is a subset of the full sample. The replicate samples were formed by grouping all respondents by stratum and then randomly selecting a half-sample from one stratum. That half-sample was given a double weight. The process was repeated for other strata until the desired number of replicates was obtained. Each replicate provides an estimate of the statistic of interest, and the variability among the replicate estimates can be used to derive an estimate of the variance of the statistic.

Depending on how the strata and pairs within strata are defined, the replication technique can also be used to estimate the separate components of variance shown in equation (3). For example, to estimate the total variance, $\sigma_W^2(y')$, the assignment of units within a stratum was made by pairing PSUs in noncertainty strata and pairing segments in certainty strata. Segments were placed in the original order of selection and assigned to each member of the pair in an alternate way. To estimate the within-PSU variances, $\sigma_W^2(y')$, the pairing was performed by segment in all strata, in both certainty and noncertainty PSUs.

The between-PSU variance was computed by subtraction as

$$\sigma_B^2(y') = \sigma^2(y') - \sigma_W^2(y').$$

Two sets of data files were created for the compositing analysis. One data set included the national sample cases in the PSUs within the 11 states. The second file combined data from the 11 state samples. The 11-state national and state sample data sets were separately weighted up to the known total population following the same weighting procedures used for the NALS file. For each of the data sets, two sets of replicates were formed to compute the total and within-PSU variances.

Compositing factors were calculated for each of the 11 states as a function of the between- and within-PSU unit variances, counts of PSUs (excluding those selected for the national sample with certainty), and respondent sample sizes. The 11 state samples were combined to ensure adequate degrees of freedom for the estimation of between-PSU variances. Compositing factors were calculated separately for national certainty PSUs and the remainder of the PSUs in the sample. As mentioned earlier, because national sampling strata and PSUs crossed state boundaries, sample weights that simply reflected the reciprocal of the probabilities of selection did not provide efficient state estimates. However, this problem affected only the estimates from noncertainty PSUs.

Because the certainty PSUs in the national sample represented only themselves (i.e., a certainty PSU constituted the entire stratum), sample cases coming from these PSUs could be directly combined with the state data. Given the difference in the reliability of estimates coming from certainty and noncertainty PSUs, separate compositing factors were computed for the two types of PSUs. Separate factors were also developed for the population subgroups for which different sampling rates were used in the national sample (i.e., blacks, Hispanics, and others). The basic form of the composite estimator is given in equations (1) and (2).

For data collected in PSUs other than those selected with certainty for the national sample,

$$V(\hat{Y}_{(nt)ik}) = \frac{m_{(nc)(nt)}\sigma^2_{(nt)bk}}{m_{(nc)(nt)i}} + \frac{n_{(nt)k}\sigma^2_{(nt)wk}}{n_{(nc)(nt)ik}} \qquad (4)$$

where

$m_{(nc)(nt)}$ = the number of national sample PSUs across the 11 states that were not selected with certainty;

$\sigma^2_{(nt)bk}$ = the national between-PSU variance for subgroup k;

$m_{(nc)(nt)i}$ = the number of national sample PSUs in state i that were not selected with certainty;

$n_{(nt)k}$ = the number of respondents in the national sample across the 11 states for subgroup k;

$\sigma^2_{(nt)wk}$ = the national within-PSU variance for subgroup k; and

$n(nc)(nt)ik$ = the number of national sample respondents not in national certainty PSUs in state i for subgroup k.

Similarly,

$$V(\hat{Y}_{(st)ik}) = \frac{m_{(nc)(st)}\sigma^2_{(st)bk}}{m_{(nc)(st)i}} + \frac{n_{(st)k}\sigma^2_{(st)wk}}{n_{(nc)(st)ik}} \qquad (5)$$

where

$m_{(nc)(st)}$ = the number of state sample PSUs across the 11 states that were not selected with certainty;

$\sigma^2_{(st)bk}$ = the state between-PSU variance for subgroup k;

$m_{(nc)(st)i}$ = the number of state sample PSUs in state i that were not selected with certainty:

$n_{(st)k}$ = the number of respondents in the state sample across the 11 states for subgroup k;

$\sigma^2_{(st)wk}$ = the state within-PSU variance for subgroup k; and

$n(nc)(st)(ik)$ = the number of state sample respondents not in national certainty PSUs in state i for subgroup k.

For data collected in PSUs selected with certainty for the national sample, the between-PSU component of the variance is equal to 0, and the formula for variance simplifies to

$$V(\hat{Y}_{(nt)ik}) = \frac{n_{(nt)k}\sigma^2_{(nt)wk}}{n_{(c)(nt)ik}} \qquad (6)$$

where

$n_{(c)(nt)ik}$ = the number of national sample respondents in national certainty PSUs in state i for subgroup k.

Similarly,

$$V(\hat{Y}_{(st)ik}) = \frac{n_{(st)k}\sigma^2_{(st)wk}}{n_{(c)(st)ik}} \qquad (7)$$

where

$n_{(c)(st)ik}$ = the number of state sample respondents in national certainty PSUs in state i for subgroup k.

Under the assumption of equal within-PSU variance for national certainty and noncertainty PSUs, data from all PSUs were combined for the estimation of this component of variance.

Note that the specific components of variance computed in equations (4), (5), (6), and (7) reflect the national and state sample designs (i.e., the oversampling of minority populations), as well as the fact that national PSUs crossed state boundaries.

Table 2 provides the estimated compositing factors for average prose proficiency and literacy levels for the state of California.

6. CONCLUSIONS

The combining of data from national and state sample surveys presents unique problems and opportunities. Although the two types of surveys often have contrasting primary domains of analysis which influence their respective sample designs, the data can be combined for increased precision of most estimates given some knowledge of the sampling variances involved. The sampling variances can be broken into components reflecting stages of selection and can either be estimated from similar, previous surveys or calculated using actual survey data. The estimated sampling variances can be used to calculate composite estimation factors, which can be embedded in the full sample and replicate weights. This paper confirmed that for NALS the precision of most estimates did increase as a result of the composite estimation.

Table 2. NALS Compositing Analysis: Optimum Compositing Factors for Average Prose Proficiency and Literacy Levels in California Sample, by Total Population, Gender, Race/Ethnicity, Education Level, Age, and Country of Birth

Demographic Subpopulations	National certainty PSU?	State Compositing Factor: Beta					
		Level 1 225 or lower	Level 2 226 to 275	Level 3 276 to 325	Level 4 326 to 375	Level 5 376 or higher	Overall Proficiency
Total Population							
Total	yes	0.2700	0.3542	0.1695	0.2901	0.5171	0.2087
	no	0.3651	0.3734	0.2392	0.3308	0.5817	0.2966
Gender							
Male	yes	0.3053	0.3134	0.2328	0.3521	0.5689	0.2387
	no	0.3389	0.3341	0.2702	0.3765	0.6337	0.2857
Female	yes	0.3599	0.2387	0.3256	0.4144	0.3022	0.2642
	no	0.4444	0.2336	0.4179	0.4974	0.3544	0.3113
Race/Ethnicity							
Black	yes	0.1758	0.1656	0.1948	0.0936	0.3683	0.0760
	no	0.2045	0.1876	0.2374	0.1056	0.4070	0.0968
Hispanic	yes	0.1960	0.2921	0.2744	0.5604	0.2753	0.1776
	no	0.2298	0.2662	0.3497	0.5404	0.2478	0.1872
Other	yes	0.4734	0.3971	0.2594	0.3358	0.6051	0.3458
	no	0.5345	0.4072	0.3057	0.3657	0.6402	0.4080
Education Level							
No HS degree	yes	0.3163	0.4314	0.3746	0.2238	0.2561	0.1669
	no	0.3711	0.5201	0.4989	0.2905	0.3212	0.2257
HS degree	yes	0.2863	0.2155	0.2959	0.3347	0.4641	0.2324
	no	0.4518	0.3178	0.4061	0.4734	0.6089	0.3986
Some college	yes	0.2505	0.3212	0.1646	0.3226	0.5557	0.2862
	no	0.3041	0.3814	0.2672	0.3524	0.5878	0.3434
College graduate	yes	0.2894	0.4628	0.4024	0.2792	0.4862	0.3973
	no	0.3227	0.5462	0.4368	0.3181	0.5556	0.4483
Age							
16 to 24 years	yes	0.2011	0.1825	0.2447	0.4839	0.2230	0.2026
	no	0.2319	0.1965	0.3080	0.5258	0.2993	0.2916
25 to 44 years	yes	0.2381	0.4498	0.2325	0.2981	0.5729	0.2411
	no	0.3016	0.4502	0.2837	0.3132	0.6196	0.2737
45 to 64 years	yes	0.2684	0.3324	0.3390	0.3481	0.2939	0.2645
	no	0.3843	0.3950	0.3845	0.4192	0.3554	0.3654
Country of Birth							
Not USA	yes	0.1557	0.2626	0.3345	0.5888	0.1894	0.1315
	no	0.1192	0.1712	0.2825	0.4653	0.1500	0.1099
USA	yes	0.3315	0.4413	0.2080	0.2882	0.5113	0.2840
	no	0.4584	0.4949	0.2929	0.3513	0.5924	0.3495